NYPD Shooting Incident Data Report

Stuart McAllister

2024-05-31

```
# Load relevant libraries for analysis project
library(tidyverse)
library(magrittr)
```

Project Step 1: Import Relevant Dataset

```
# Download and read in the csv file provided by the NYC OpenData
# website related to historic shooting incidents in New York City
# from 2006-2023

url_NYPD <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shooting_data <- read_csv(url_NYPD)
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Take an initial view of imported dataset
shooting_data
```

```
## # A tibble: 28,562 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
##  1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
##  2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
##  3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
##  4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
##  5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
##  6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
##  7    229224142 06/07/2021 19:55      QUEENS    <NA>                   113
##  8    231246224 07/22/2021 01:47      BROOKLYN  <NA>                    77
```

```
##  9    228559720 05/22/2021 18:39      BRONX     <NA>                    48
## 10    238210279 12/22/2021 23:17      BRONX     <NA>                    49
## # i 28,552 more rows
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

```
# Create a summary view of the dataset to observe for significant missing
# data that could affect later anaylsis
summary(is.na(shooting_data))
```

```
##  INCIDENT_KEY    OCCUR_DATE      OCCUR_TIME         BORO
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:28562     FALSE:28562     FALSE:28562     FALSE:28562
##
##  LOC_OF_OCCUR_DESC  PRECINCT       JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Mode :logical     Mode :logical   Mode :logical     Mode :logical
##  FALSE:2966        FALSE:28562     FALSE:28560       FALSE:2966
##  TRUE :25596                       TRUE :2           TRUE :25596
##  LOCATION_DESC   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX
##  Mode :logical   Mode :logical           Mode :logical   Mode :logical
##  FALSE:13585     FALSE:28562             FALSE:19218     FALSE:19252
##  TRUE :14977                             TRUE :9344      TRUE :9310
##  PERP_RACE       VIC_AGE_GROUP    VIC_SEX         VIC_RACE
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:19252     FALSE:28562     FALSE:28562     FALSE:28562
##  TRUE :9310
##  X_COORD_CD      Y_COORD_CD       Latitude       Longitude
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:28562     FALSE:28562     FALSE:28503     FALSE:28503
##                                  TRUE :59        TRUE :59
##   Lon_Lat
##  Mode :logical
##  FALSE:28503
##  TRUE :59
```

Project Step 2: Tidy and Transform Data

Using the initial view of the dataset, along with summary of NA values to indicate missing data there are a few things that can be done to tidy the dataset. The goal is to put it in a form that allows for ease of use and analysis. The following changes will be made:

1. Currently the OCCUR_DATE observations are in a character format which makes it difficult for using to analyze or create visualizations in this project. I will use the lubridate parse function 'mdy' to convert these to a date format.
2. Next there are a number of columns of data that we will not use for this analysis project. Apart from BORO I will not use any of the location or location description information. It is worth noting that much of the descriptive location data is missing from the dataset anyway, which makes it less useful for comparison.

3. It is also clear from the NA data summary that a large percentage of the perpetrator data is missing from the dataset. This makes sense in that it is more likely to recognize an incident from its 'aftermath' (ie. the victim) than from the perpetrator. For this reason the perpetrator data will be removed, and we will focus on the incidents from the victim's perspective.
4. The VIC_AGE_GROUP is currently a character type, but is divided into specific categories and as such I will convert it to a factor type. At the same time there appears to be one entry for VIC_AGE_GROUP that does not fit with the other categories of this data, and so I will lump this one outlier with the other unknown values using fct_lump.
5. Each incident is registered to a specific date in the dataset, but there is some analysis that requires consolidating the incidents on an annual basis. I will create a new column which extracts the 'year' value only from each date.

```
shooting_data <- shooting_data %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    subset(select = -c(LOC_OF_OCCUR_DESC:LOCATION_DESC,
                       PERP_AGE_GROUP:PERP_RACE, X_COORD_CD:Lon_Lat)) %>%
    filter(!is.na(VIC_AGE_GROUP)) %>%
    mutate(VIC_AGE_GROUP = fct_lump(VIC_AGE_GROUP, n = 5))

# Create a new column called YEAR which extracts only the year value from
# each incident's date
shooting_data$YEAR <- format(shooting_data$OCCUR_DATE, format = "%Y")

# Print another view of the dataset after modifications
shooting_data
```

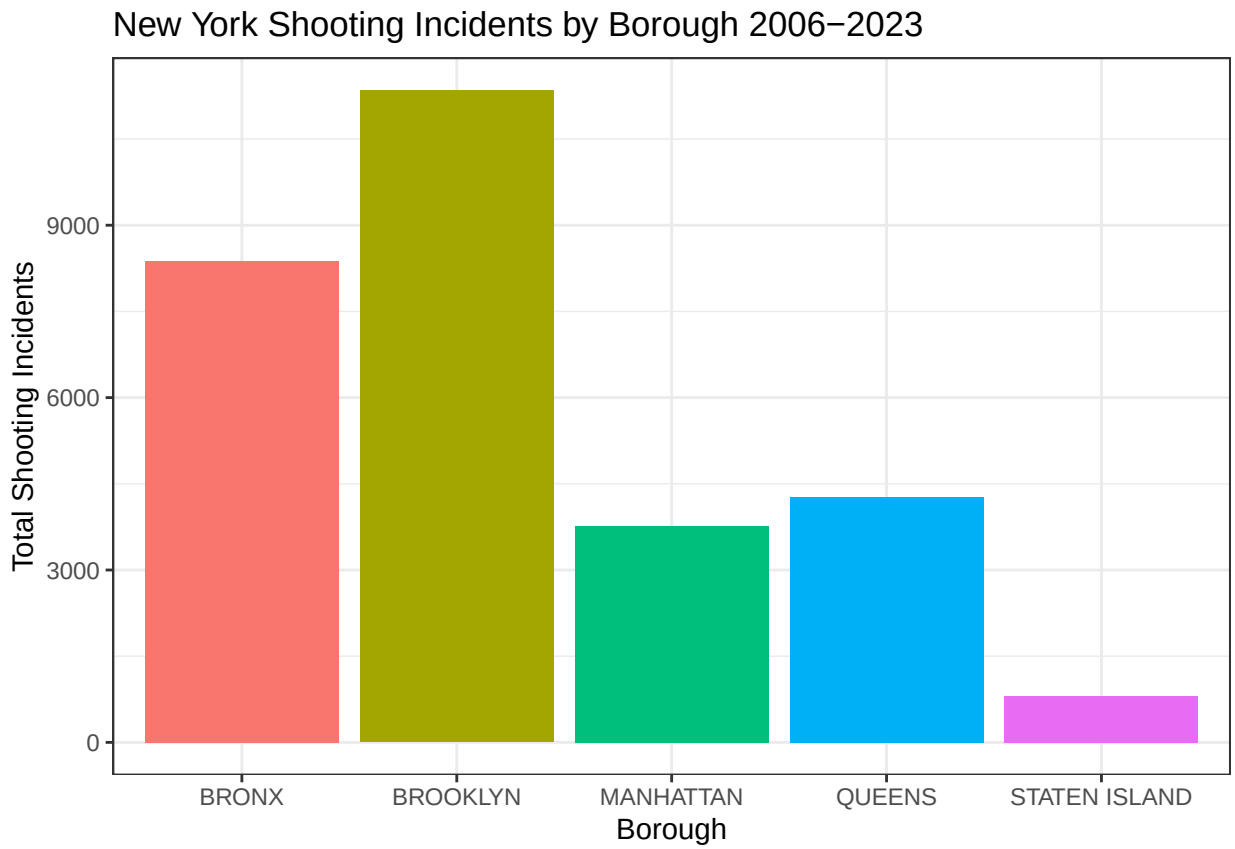
```
## # A tibble: 28,562 x 9
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO  STATISTICAL_MURDER_F~1 VIC_AGE_GROUP
##           <dbl> <date>     <time>     <chr> <lgl>                  <fct>
##  1    244608249 2022-05-05 00:10      MANH~ TRUE                   25-44
##  2    247542571 2022-07-04 22:20      BRONX TRUE                   18-24
##  3     84967535 2012-05-27 19:35      QUEE~ FALSE                  18-24
##  4    202853370 2019-09-24 21:00      BRONX FALSE                  25-44
##  5     27078636 2007-02-25 21:00      BROO~ FALSE                  25-44
##  6    230311078 2021-07-01 23:07      MANH~ FALSE                  25-44
##  7    229224142 2021-06-07 19:55      QUEE~ TRUE                   45-64
##  8    231246224 2021-07-22 01:47      BROO~ FALSE                  25-44
##  9    228559720 2021-05-22 18:39      BRONX FALSE                  18-24
## 10    238210279 2021-12-22 23:17      BRONX TRUE                   25-44
## # i 28,552 more rows
## # i abbreviated name: 1: STATISTICAL_MURDER_FLAG
## # i 3 more variables: VIC_SEX <chr>, VIC_RACE <chr>, YEAR <chr>
```

Project Step 3: Visualization and Analysis

To begin with, I would like to see if there are differences in the number of incidents over the years 2006-2023, based on geographical location within the city of New York. I will create a bar chart of total number of registered incidents for each of the 5 boroughs of New York from the dataset.

```
# Create a bar chart of total shooting incidents per borough for 2006-2023.
shooting_data %>%
    ggplot(aes(x = BORO)) +
```

```
    geom_bar(aes(fill = BORO), show.legend = FALSE) +
    labs(title = 'New York Shooting Incidents by Borough 2006-2023', x =
           'Borough', y = 'Total Shooting Incidents') +
    theme_bw()
```



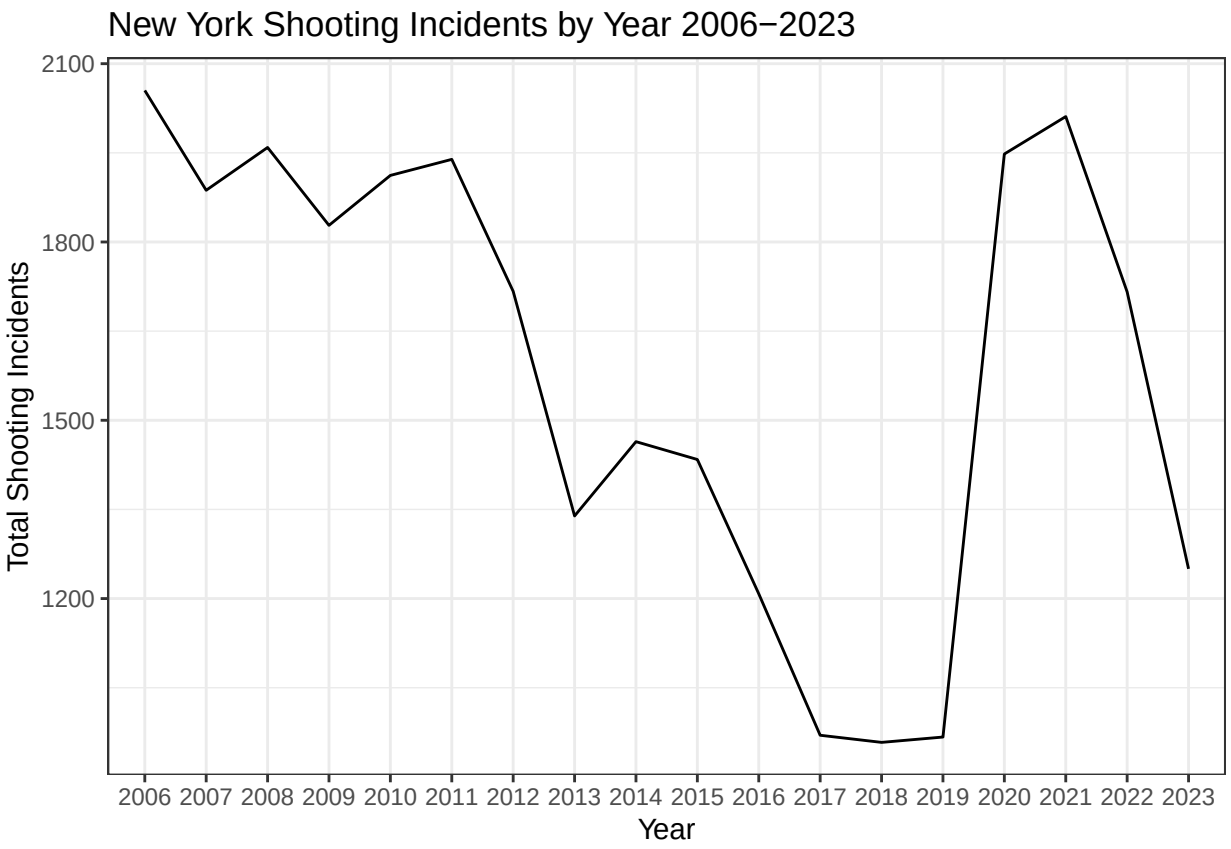
It appears that there are significantly more shootings in Brooklyn followed by the Bronx based on the data presented. In comparison, Staten Island has seen a total of less than 10% of the total experienced in Brooklyn.

Further to this I would like to see if there are general trends based over time for shooting incidents in New York City. I will use the total incidents grouped by year to create a line chart that shows the tendencies over the timeframe of the dataset.

```
# Group the shooting incidents by year and summarise with the variable
# total_shootings per year
incidents_per_year <- shooting_data %>%
    group_by(YEAR) %>%
    summarise(total_shootings = n())

# Create a line chart showing the changes in shooting incidents over
# the years of the dataset
    ggplot(data = incidents_per_year,
           aes(x = YEAR, y = total_shootings, group = 1)) +
        geom_line() +
        labs(title = 'New York Shooting Incidents by Year 2006-2023',
                    x = 'Year', y = 'Total Shooting Incidents') +
```

```
        theme_bw()
```



There appears to be a fairly steady decline of shooting incidents from the beginning of the dataset in 2006 until 2019. Then we can see a dramatic increase in shootings from 2019 until a peak amount in 2021. It appears that shootings more than doubled in than two year period, after which we see another period of decline.

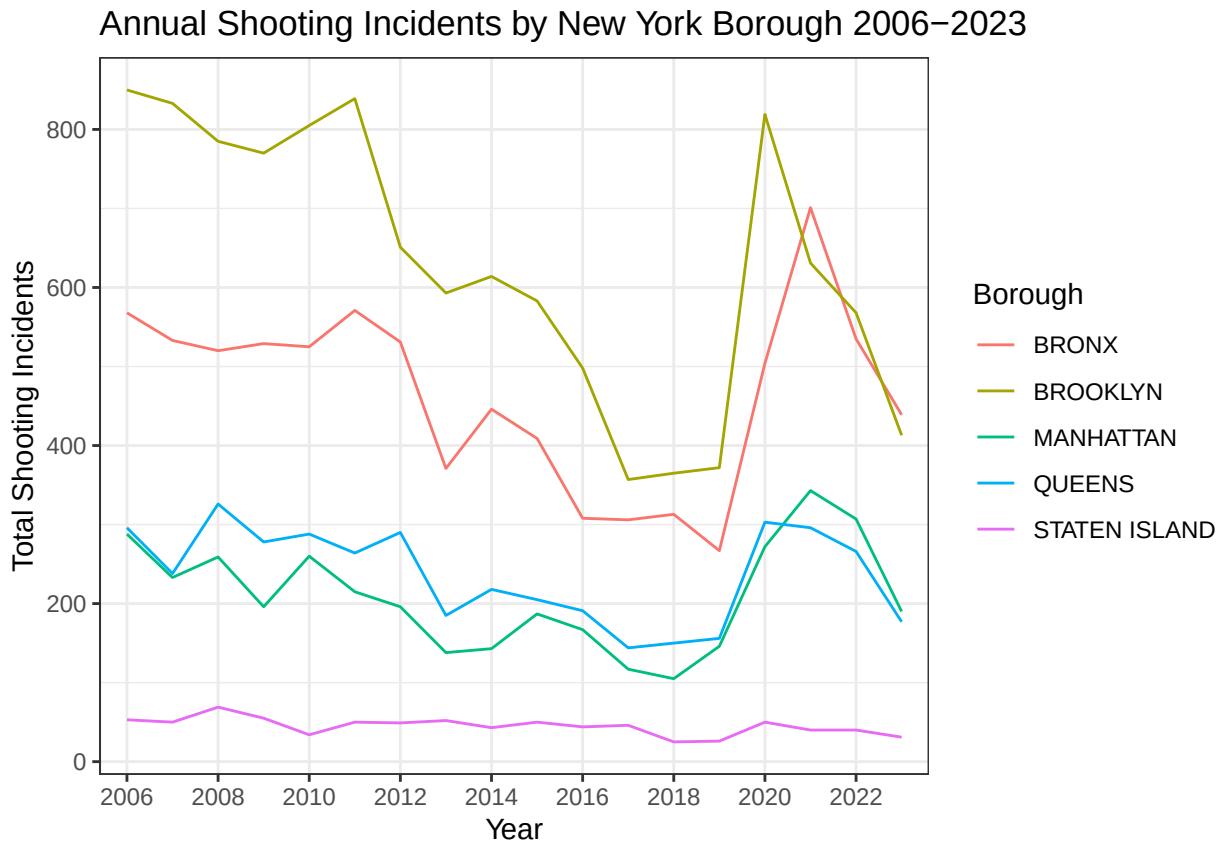
This visualization leads me to want to further understand this trend in 2020. Was it related to the chaos created in the entire world due to the COVID-19 pandemic? Where there other societal pressures, eg. race relations, economic stability, government policies, etc? Was there less police presence due to involvement as first responders for pandemic situations? Are these trends similar in other large US or World cities? All of these could be further investigated, using related data to compare trends and look for correlations.

Combining the previous two visualizations, I would like to see if this spike in 2020 is correlated with different geographic locations in New York City. I will group the incidents by year and by borough, and create a line chart to show the incident trends over time in each of the 5 boroughs.

```
# Group the shooting dataset by both year and borough and summarize to see
# total shootings per borough
incidents_by_year_boro <-shooting_data %>%
        group_by(YEAR, BORO) %>%
        summarise(total_shootings = n())
```

```
## `summarise()` has grouped output by 'YEAR'. You can override using the
## `.groups` argument.
```

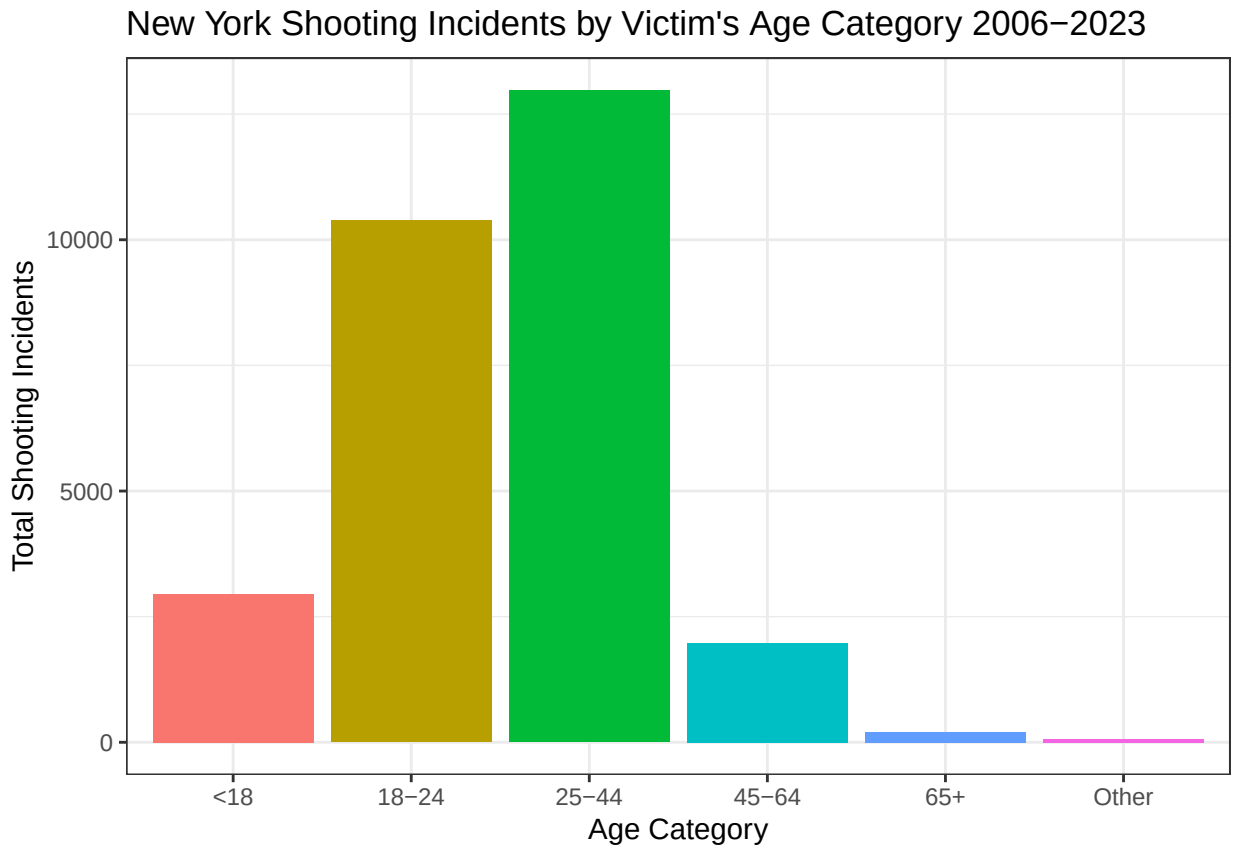
```r
# Create a line chart based on annual shootings per borough over time
        ggplot(incidents_by_year_boro,
               aes(x = YEAR, y = total_shootings, group = BORO)) +
          geom_line(aes(color = BORO)) +
# Reduce the number of labels on the x-axis for clarity
          scale_x_discrete(breaks = seq(2006, 2024, 2)) +
          labs(title = "Annual Shooting Incidents by New York Borough 2006-2023",
               x = 'Year', y = 'Total Shooting Incidents', color = 'Borough') +
          theme_bw()
```



From the visualization there appears to be a fairly noticeable spike in shooting incidents in at least 4 of the 5 boroughs in 2020, which lines up with the overall yearly trends for the entire city. A statistical analysis could verify these results, but the trends seem to be apparent. On the other hand Staten Island seems to remain relatively flat. Are there characteristics of this borough that make it less propense to shooting incidents? Where there factors in 2020 and 2021 that were more impactful in the other boroughs? These are areas that could be studied with further data.

Finally, I would like to look at the age demographics of the victims involved in shooting incidents in the dataset. I will create a bar chart divided by age group category to see if there are any apparent trends or characteristics.

```r
# Create a bar chart of incidents divided by age category
      ggplot(shooting_data, aes(x=VIC_AGE_GROUP)) +
          geom_bar(aes(fill = VIC_AGE_GROUP), show.legend = FALSE) +
          labs(title = 'New York Shooting Incidents by Victim\'s Age Category 2006-2023', x = 'Age Categ
          theme_bw()
```



A significant amount of shooting victims can be seen in the age ranges of 18-25 and 25-44, as compared to age groups above and below this amount. Further investigation could help to explain these trends, as well as if they relate to other aspects identified - geographic disparity, spike in incidents around 2020-2021. Are there other demographic or societal factors that contribute to the large percentage of victims being in this age range? Could this be used to speak to causal aspects?

Project Step 4: Conclusion and Bias Identification

The NYPD Historical Shooting Incident dataset provides a glimpse into the aspects of gun violence that play a role in many, if not most of the world's large metropolitan areas. Because of the population size of New York City, it can be easier to see trends and patterns, simply due to the quantity of incidents. From initial analysis, there seem to be patterns in New York's gun violence related to demographics, geography, and time.

As with most investigations, some questions may be answered immediately, but often the analysis simply leads to further opportunities to dive deeper. From this initial analysis I can see the following as further areas of investigation:

1. How does New York City's shooting incident patterns compare with those of other large US cities and/or comparable world cities?
2. What are the factors that influenced a rapid increase in shooting incidents in 2020-2021 compared with a general downward trend seen in the dataset?
3. Why are young adults more commonly the victims of shooting incidents in New York City?
4. What are the factors that influence different shooting incident patterns within the geographical areas of New York City, and how can this information be useful for identifying potential similar areas in other urban centers?

Bias Identification

We all grow up in a context which shapes the way that we interpret the world around us is. It is important to identify some of the biases that might influence the analysis and interpretation of the data that we are given to work with. Here are some of the biases that I identify in this project, and how i tried to deal with them:

1. I purposely left out racial characteristics when sorting and visualizing the data. There is a general preconception that a correlation exists between race and crime and/or gun violence. This can be a divisive topic and as such I wanted to look at the data without the influence of race. In doing so it is possible that I might be filtering some important insights into the questions of difference in location and age demographic. Any further analysis would need to keep this in mind so as not to lead the investigation based on preconceived notions.
2. There is a significant amount of data missing, especially in terms of the perpetrators of the shooting incidents. I decided to filter this out to focus on the victims, but there could be some important trends or insights to be learned even with the limited perpetrator information. If handled carefully, this information could be incorporated into further study.
3. I have a limited knowledge of New York City, which can both complicate the analysis of the dataset and also help to keep a neutral perspective when approaching the observed data. Some prior knowledge of the information being analyzed could be useful to create a path for research, as long as it doesn't create preconceptions.
4. I am working under the assumption that the people that are recording this data are doing it without their own biases. On a large sample like this, that may be a fair assumption, but that doesn't mean that there aren't influences of bias in the data itself in how crimes are categorized or classified.

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/El_Salvador
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] magrittr_2.0.3  lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
##  [9] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.5      highr_0.10        crayon_1.5.2
##  [5] compiler_4.4.0    tidyselect_1.2.1  parallel_4.4.0    scales_1.3.0
##  [9] yaml_2.3.8        fastmap_1.1.1     R6_2.5.1          labeling_0.4.3
```

```
## [13] generics_0.1.3    curl_5.2.1        knitr_1.46        munsell_0.5.1
## [17] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4
## [21] stringi_1.8.4     xfun_0.43         bit64_4.0.5       timechange_0.3.0
## [25] cli_3.6.2         withr_3.0.0       digest_0.6.35     grid_4.4.0
## [29] vroom_1.6.5       rstudioapi_0.16.0 hms_1.1.3         lifecycle_1.0.4
## [33] vctrs_0.6.5       evaluate_0.23     glue_1.7.0        farver_2.1.1
## [37] fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.26    tools_4.4.0
## [41] pkgconfig_2.0.3   htmltools_0.5.8.1
```